Towards a semantic model to enhance knowledge sharing and discovery in organic chemistry

Valentina Dragos, Adeline Nazarenko

TOWARDS A SEMANTIC MODEL TO ENHANCE KNOWLEDGE SHARING AND DISCOVERY IN ORGANIC CHEMISTRY

Valentina Dragos
Centre de Recherche des Cordeliers
Université Pierre et Marie Curie - Paris6 UMR_S 872 ; Université Paris Descartes, UMR_S 872 ; INSERM, U872
5, Rue de l'Ecole de Médecine
Paris, F-75006 France

Adeline Nazarenko
LIPN – Université Paris 13 & CNRS
99, avenue J.B. Clément
Villetaneuse, F-93430 France

ABSTRACT

This paper presents the project of an electronic encyclopaedia of organic Chemistry. The goal of the EnCOrE (Encyclopédie de la Chimie Organique Electronique) encyclopaedia is twofold: first, it aims at enhancing knowledge sharing in organic Chemistry, by providing a unified access to the increasing amount of domain resources; second, it aims at improving knowledge discovery in the field of organic Chemistry, by revealing unknown connections between those resources. EnCOrE encyclopaedia is designed for a broad category of users such as students, researchers, and teachers.
The paper introduces our vision of the EnCOrE encyclopaedia as an information system. It presents its architecture and describes the semantic model sustaining this architecture. Critical points and challenges related to the EnCOrE project are also discussed.

KEYWORDS

Information system, information management, semantic model, encyclopaedia, organic Chemistry.

1. ENCORE: AN OPEN AND CONTROLLED ENCYCLOPAEDIA OF ORGANIC CHEMISTRY

Organic Chemistry is a dynamic field, strongly connected to several other research fields such as Medicine, Biology or Physics. According to (Krief. *et al*., 2003), chemists acquire, represent and transfer their knowledge through chemical experiments (Lowenthal *et al.,* 1990, Harwood *et al*., 1983), chemical equations and documents (manuals, scientific articles) related to bibliographical references. The fast domain evolution leads to an increasing amount of heterogeneous information (scientific articles, chemical equations, or bibliographical references). The use of information technologies to handle this large amount of information becomes obvious to Chemistry practitioners. Therefore, several artefacts have already been created in the form of electronic databases and document collections (*e.g.* Beilstein[1]). Recently, three projects have emerged that improve considerably the field of chemical information. PubChem[2] is a free service of free databases incorporated into a unique information retrieval system. The databases contain compounds structures and provide information of their biological property. DiscoveryGate[3] is a service providing access

[1] http://www.beilstein.com
[2] http://pubchem.ncbi.nlm.nih.gov
[3] https://www.discoverygate.com

to more than 20 databases through a unified platform. The databases mainly contain experimental data, derived from different sources such as scientific journals, patents, or chemical catalogs. Chemistry Central[4] is a service providing access to databases of scientific articles of Chemistry and related fields (ChemCentral).
New developments in chemical information systems are also made possible by recent advances of molecular informatics which aim, for instance, at designing an XML language for Chemistry (Murray-Rust *et al*., 2003), at normalising chemical name nomenclature (Pîrkola, 2008) or at creating large repositories of experimental data (Morgan, 2007).
However, in spite of those advances, Chemistry remains slower than other disciplines, like Biology or Physics, in adopting information technology solutions to improve the organization of chemical information and to provide efficient access to this information. Nowadays systems suffer from several major limitations: they provide access to homogenous resources and it is not possible to retrieve, in a unified manner, chemical equations and scientific articles related to the same compound, for instance; they ignore the user profile and are therefore unable to provide information at an appropriate level, because a young student and a researcher may have different information needs; they miss advanced functionalities to support knowledge discovery beyond access to large amount of information. Despite similarities, Chemistry differs from other domains like Biology in the way knowledge is encoded. Chemists have developed formal and graphic knowledge representation tool and scientific texts do not have the same importance as in Biology. Text mining technologies have therefore a different role too play in Chemistry.
By designing EnCOrE, our goal is to propose a new organization for chemical information, coping with limitations of actual systems. EnCOrE will be able to give new insights, to reveal unknown connections and to enhance knowledge sharing and discovery in organic Chemistry field. By using domain knowledge, our system will be closer to the Telemakus system (Fuller *et al*., 2004), a knowledge-based system designed to enhance retrieval and review of scientific research reports across a domain. From a general point of view, EnCOrE could be described as an information system allowing different users to access various resources and to create new ones.
Resources of EnCOrE (henceforth, *information sources*) are either internal – created by EnCOrE users –or external – stocked in external databases or repositories. The internal resources are encyclopaedia articles and the chemical dictionary. They cover various topics, such as domain history, important scientific discoveries, pioneers of Chemistry, and are created by Chemistry experts who ensure the quality of information. Internal resources are created and updated during EnCOrE exploitation. External resources are chemical equations, compounds structures, images, scientific articles and bibliographical references. They are stocked in external databases or repositories which often require subscription to be accessible and which existence and evolution are not related to EnCOrE, which only provides a unified access to those resources.
Following sections will introduce the architecture of EnCOrE and will describe its user categories.

2. A SERVICE ORIENTED ARCHITECTURE

We propose a service-oriented architecture in order to develop and implement the EnCOrE encyclopaedia. Therefore, the EnCOrE architecture consists of a number of services allowing to access information sources, to create new knowledge pieces and to update the existing ones as well as interfaces allowing different users to select the appropriate services, according to the task they want to achieve.
From a service oriented point of view, EnCOrE can be defined as an open – resources of EnCOrE will be available via internet – and controlled – new resources will be created only by a particular category of users[5] – encyclopaedia of organic Chemistry. By making this choice, our goal is to provide valid information to a broad category of users. In this respect, EnCOrE is closer to Scholarpedia[6], in which information is added by prominent authorities of various research fields, than to Wikipedia[7], in which any user can add information. However, extending Scholarpedia is not sufficient, as our goal is to provide, beyond information access and retrieval, advanced functionalities such as new knowledge discovery or emergence.

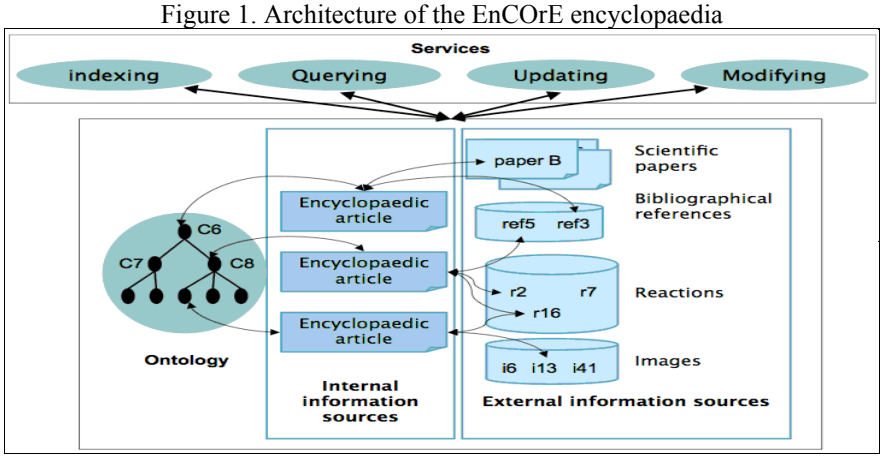
[4] http://www.univarusa.com/chemcentral.htm
[5] This is a requirement of the chemists that promote the EnCOrE project (Krief *et al*. 2003).
[6] http://www.scholarpedia.org/
[7] http://fr.wikipedia.org/

We propose a semantic model to ensure interoperability and communication within the heterogeneous collection of EnCOrE information sources. This model uses an ontology and relies on a double indexing system. An ontology is defined by (Gruber, 1993) as a formal and explicit formalization of a shared conceptualization. Ontology models domain specific concepts and relations holding between them. Various relations can be made explicit. Among them, the most important is the generalisation relation that organises the set of concepts in a hierarchy. Although building the relevant ontology is a critical aspect of the EnCOrE implementation - we plan to use Collaborative Protégé (Tudorache *et al.,* 2008) and text-based knowledge acquisition tools (Cimiano, 2006)- this point is not addressed in this paper.

A double indexing system is also at the core of our semantic model. A first indexing system assigns an ontology concept to each encyclopaedia article. Therefore, internal resources of EnCOrE are connected to each other through the ontology conceptual structure. A second indexing system connects internal and external information sources. It assigns external entities of external databases (*e.g.* chemical reactions, images) to each encyclopaedia article. The originality of our semantic model is that external entities are referred in both intentional (a list of queries allowing to retrieve entities) and extensional (list of entities) manners, which is important to keep the whole system consistent and up-to-date.

Ontological concepts and external entities represent two different types of metadata describing the encyclopedia articles. This articulation between ontology concepts and external entities defines a semantic model sustaining the EnCOrE encyclopedia.

Figure 1. Architecture of the EnCOrE encyclopaedia



This semantic model creates indirect links between external resources, see Figure 1. This way, EnCOrE provides unified access to external resources but it is also able to go beyond the flat structure of those information sources and reveal unknown links between them. Therefore, the semantic model enhances knowledge emergence or discovery in organic Chemistry, thanks to reasoning capabilities of the ontology. For instance, one could analyse the similarities between entities describing encyclopaedia articles indexed by a given ontology concept, the consistency of the set of articles referring to a given reactions, the co-citations between two specific substances. However, we do not propose or anticipate any knowledge discovery procedure, as our goal is to offer a rich semantic structure supporting a wide range of exploration strategies.

3. USERS OF THE ENCORE ENCYCLOPAEDIA

Knowledge sharing and discovery is achieved thanks to different users of the EnCOrE encyclopaedia. Each user has a particular profile, according to which it can achieve particular tasks. We have identified four categories of EnCOrE users. G*rand public* users are allowed to navigate through the collection of information sources through querying services. These services include information retrieval functionalities and exploit the EnCOrE underlying semantic model. Since the user's information need is often not clearly expressed, they also help the users to better precise his information need. *Author* users are allowed to add new encyclopaedia articles. EnCOrE will include an annotation or indexing service that helps authors to annotate their article with respect to the semantic model of EnCOrE. These services exploit text mining methods to propose to the author a set of concepts as indexing keys for a given article. If the author considers the set as irrelevant for

the article, it can trigger an alert and indicate his own index suggestions. The result of such an indexing procedure could be either a new indexed encyclopaedia article, properly integrated in a heterogeneous collection, or a stand-by encyclopaedia article and several index suggestions. *Knowledge manager* users behave as an advisory board that works collaboratively to create and maintain EnCOrE knowledge. From a practical point of view, the knowledge manager ensures the ontology evolution, which is a challenging task. Two mechanisms are envisioned to trigger the ontology update. The first one deals with the coverage of the topic ontology, which must be updated if it does not adequately describe the information sources (either too many documents are indexed by the same concepts, or some concepts do not index any document). The second mechanism exploits index suggestions that authors make while indexing their papers in addition to existing ontological concepts. These new descriptors can be turned into concepts by the knowledge manager at a later stage. By using those mechanisms, the ontology reflects the domain evolution and encyclopaedia articles are always connected to last minute Chemistry advances. Knowledge managers exploit services to create knowledge, detect inconsistencies in it and update it. In addition, a mechanism is provided to periodically check the consistency of the extensional and intentional references to database entities assigned to encyclopaedia articles. If both references do not match, it indicates that the refereed database has been updated (the set of records corresponding to the database query has changed) and that the metadata assigned to the encyclopaedia must be revised. *Information system manager* is a particular user in charge with system administration (adding new authors, for instance). However, those different profiles are not exclusives, and the same user can have many roles. Different interactions between users and the EnCOrE encyclopaedia will ensure knowledge sharing and emergence in organic Chemistry.

4. CONCLUSION

This paper presents an ongoing work aiming to develop EnCOrE, an electronic encyclopaedia of organic Chemistry. We describe EnCOrE from an information system point of view and introduce its architecture. This architecture is sustained by a semantic model whose advantages are also discussed. Future work will focus on the design and implementation of services which should help users to navigate through a large collection of information sources, authors to adequately index their documents, and knowledge managers to create and update the semantic model underlying EnCOrE.

REFERENCES

Cimiano, P. 2006, *Ontology Learning and Population from Text*, Springer, New York, USA.

Fuller, S. et al., 2004, A knowledge-based system to enhance scientific discovery: Telemakus, *Biomedical Digital Libraries, Biomedical Digital Libraries*, Vol. 1, No. 2.

Gruber, T., 1993, A translation approach to portable ontology specifications, *Knowledge Acquisition for Knowledge-Based Systems*, Vol. 5, No. 2, pp. 199-220.

Harwood, L.M. et al., 1983 *Experimental Chemistry: Principles and Practice*, Backwell Scientific Publications, Oxford, England.

Krief, A. et al., 2003, EnCOrE (Encyclopédie de chimie Organique Electronique): an Original Way to Represent and Transfer Knowledge from Freshmen to Researchers in Organic Chemistry, 2003, *Proceedings of the 3rd International LeGE-WG Workshop: GRID Infrastructure to Support Future Technology Enhanced Learning*, Berlin, Germany.

Lowenthal, H.J.E, 1990, *A guide for the perplexed Organic Experimentalist*, John Wiley and Sons, New York, USA

Morgan, P., 2007, Facilitating the Deposit of Experimental Chemistry Data in Institutional Repositories: Project SPECTRa, *Proceedings of the 28th IATUL annual conference: "Global Access to Science - Scientific Publishing for the Future*, Stockholm, Sweden.

Murray-Rust, P. et al., 2003, Chemical Markup, XML and the World-Wide Web. 4. CML Schema, *Journal of Chemical Information and Computer Science*, Vol. 43, No. 3, pp. 757-772.

Pîrkola, A., 2008, Extracting variant forms of chemical names for information retrieval, *Information Research*, Vol. 13 No. 3.

Tudorache, T. et al., 2008, Collaborative Protege: Enabling Community-based Authoring of Ontologies*, Proceedings of the 7th International Semantic Web Conference,* Karlsruhe, Germany.